3 「情報」応用の開拓

～全世界のWeb情報アーカイブ構築への挑戦～

山名早人†

キーワード　検索エンジン，Webアーカイブ，分散処理

1．まえがき

「計算」の応用の開拓に続き，「情報」の応用の開拓では，まず，インタネット上に存在する膨大な情報について，その推定規模とWWW（World Wide Web）が誕生してからの動向を紹介する．次に，いわゆる検索エンジンで探すことのできる情報量が，全世界からWWWを使って一般に発信されている情報の40%以下であり，検索エンジンを使ってもすべての情報を検索することができないことを示す．

そして，このような背景のもと，全世界からWWWによって発信される情報をさらに有効活用するための試みを二つ紹介する．最初は，筆者らがこれまでに実施してきた日本国内のWebページを24時間以内で収集することを目指した分散収集実験である．二つ目は，2003年度より開始した全世界のWeb情報アーカイブ構築を目指すプロジェクトである．

2．世界と日本のWebページ数推計

本節では，これまでのインタネットの発展，およびWebページ数の増加について，統計情報を基に解説する．

2.1　インタネットの発展

図1および図2に示すように，インタネットに接続されるコンピュータ台数*1は，毎年増加を続けている．1994年にWWWサーバ数が前年比16倍で増加した背景には，Webブラウザとしてそれまで公開されていたMosaicに加えてNetscapeが登場し，WWWが一般的なものになったことが大きな理由となっている．その後も2000年までは，WWWサーバ数は，年率2倍以上のスピードで指数関数的に増加を続けており，1998年から2001年の平均でみるとコンピュータ台数，WWWサーバ数は，それぞれ年率約1.5倍，2倍の増加となっている．しかし，2002年は，この傾向に変化がみられ，ドメイン数およびWWWサーバ数が前年比でそれぞれ8%減，2%減となった．これは，主として世界的なIT不況による影響によるものと考えられる．なお，図2中には示されていないが，2003年に入ってからのWWWサーバ数は，再び増加に転じており，今後もさらなる増加が予想される．

次に，インタネットに接続するコンピュータ台数に占めるWWWサーバ台数の推移を図3に示す．図3から，インタ

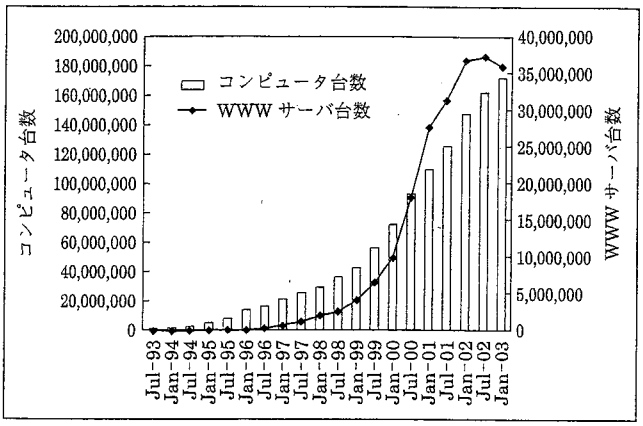


図1　インタネットに接続するコンピュータ台数とWebサーバ数の推移*2

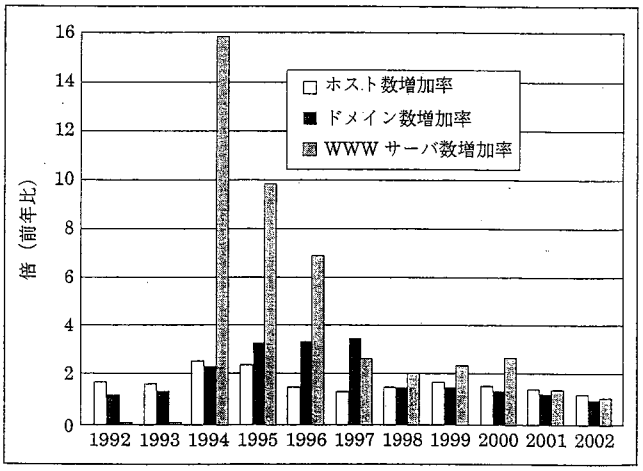


図2　インタネットに接続するコンピュータ台数，ドメイン数，WWWサーバ数の増加率*1

*1　グローバルIPを持つコンピュータ台数を示す．このため，組織内などでプライベートIPを持つコンピュータ台数はカウントされていない．

*2　Internet Software Consortium（http://www.isc.org/）のInternet Domain Survey，およびNetcraft社（http://www.netcraft.co.uk/）のWWW Server Surveyの公開データを基に作成．

†早稲田大学 理工学部 コンピュータ・ネットワーク工学科
"Exploitation of Informational Applications -Toward the Global Web Information Archive-" by Hayato Yamana (Waseda Univ., Tokyo)

映像情報メディア学会誌 Vol. 57, No. 12, pp. 1632～1637 (2003)

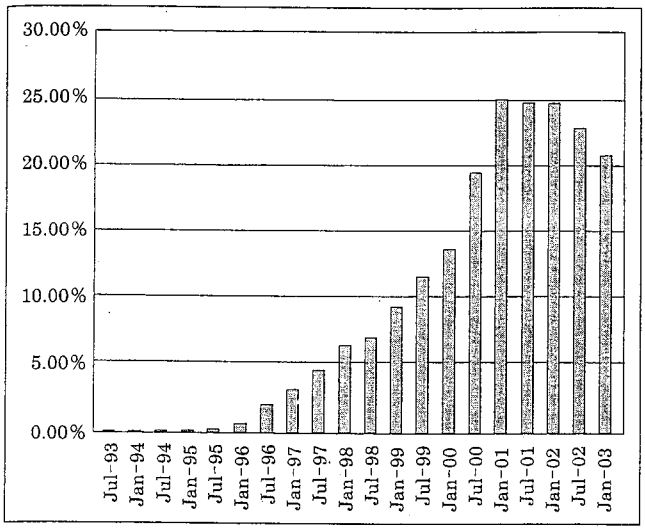


図3　インタネットに接続するコンピュータ台数に占めるWWWサーバ台数の割合*1

順位	分類	全コンピュータ台数に占める割合	推定Webページ数(億)
1	Networks	36.1%	28.1
2	Commercial	23.6%	18.4
3	日本	5.4%	4.2
4	Educational	4.3%	3.4
5	イタリア	2.3%	1.8
6	カナダ	1.7%	1.4
7	ドイツ	1.7%	1.3
8	イギリス	1.5%	1.2
8	オーストラリア	1.5%	1.2
10	スイス	1.4%	1.1
	合　計	100%	78.0

表1　ドメイン別Webページ数推計（2003年8月）

ネットに接続されるコンピュータの4～5台に1台は何らかの形で情報を発信していることがわかる．このように，インタネットは，電子メールを使ったり，WWWを使って情報を入手したりするためのものから情報を発信するためのものに変化してきている．

2.2　Webページ数の世界分布

次に，全世界に存在するWebページ数の分布に着目する．表1に，2003年8月時点でのトップレベルドメイン毎のWebページ数（テキストページ総数）の推計を示す．

NEC北米研究所のLawrenceらが提案した手法[1]に基づき，2003年8月時点での全世界のWebページ数を78億ページと推定し，各ドメイン内のコンピュータ台数にWWWサーバ台数が比例すると仮定し*3，トップレベルドメイン毎のWebページ数を算出した．なお，各ドメイン内のコンピュータ台数は，Internet Domain Survey[2]のデータに基づいている．

表1より，JPドメイン内には約4.2億のWebページが存在しており，米国のNETドメイン，COMドメインに続いて3位であることがわかる．このように，日本からは非常に多くの情報が発信されている*4．

3．商用検索エンジンのデータ規模

全Webページ数（推定）に対する商用検索エンジンのカバー率と全世界のWebページ総数の推移を図4に示す．なお，カバー率とは，全Webページ数に対する検索可能なWebページ数の割合である．

2000年からカバー率が上昇に転じているのは，1998年に誕生したGoogle[3]（http://www.google.com/）が検索対象となるWebページ数を増加させてきたことに起因する．Googleは，2000年1月時点で2.5億，7月に3.6億，2001年1月に6億，7月に10億，さらに，2002年1月に20億，7月に26億，12月に30億，2003年8月に33億と着実に検索対象となるWebページ数を増大させている．

2003年8月21日には，Overtune社が運営するデータ規模で世界第二位のalltherweb（http://www.alltherweb.com/）が，それまでの21億ページから31.5億へと検索対象Webページ数を増加させ，収集ページ数において第一位となったが，その2日後の8月23日には，Googleが検索対象Webページ数を33億として巻き返した．このことからもわかるように，商用検索エンジンの世界では熾烈な競争が繰り広げられている．

しかし，一方で，検索対象となるWebページ数を増加させるためには，膨大な設備投資が必要であり，検索エンジンが検索対象とするWebページは，全世界のWebページの

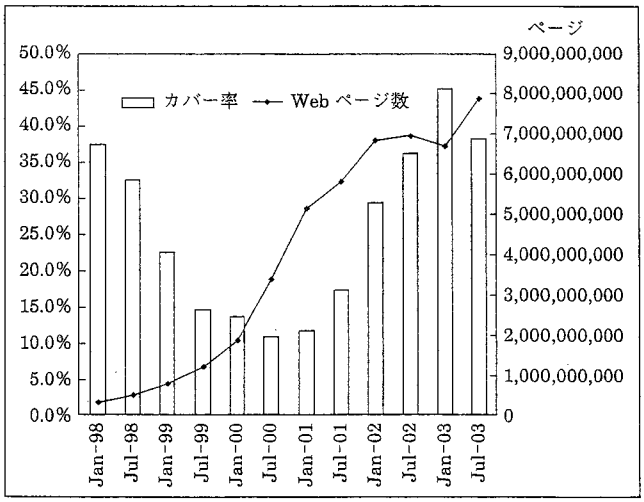


図4　推定Webページ総数に対する商用検索エンジンのデータカバー率の推移

*3　必ずしもコンピュータ台数とWWWサーバ台数は比例しない．また，正確な値を出すためには，トップレベルドメイン別のWWWサーバ数が必要となるが，推計を含めて公表されている数値は存在しない．

*4　JPドメイン内から発信されるWebページ数の推計であり，COMドメイン等の海外のドメインを使って日本語で発信されている情報は含んでいない．

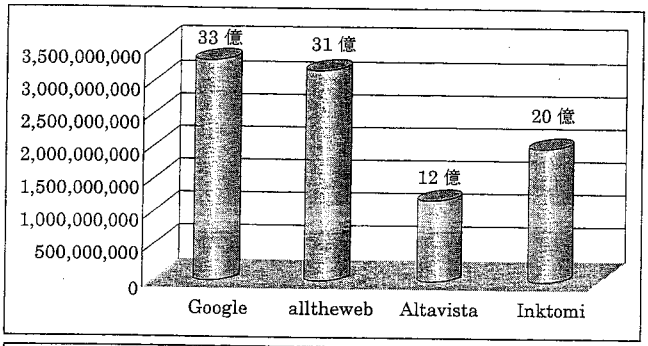


図5 商用検索エンジンのデータ規模（2003年8月）

40%程度に留まる．このため，このような競争がなくなれば，図4の2000年までの傾向に示されるように，カバー率はさらに低下するものと考えられる．

図5は，2003年8月時点における主要な検索エンジンのデータ規模を示す．2002年までは，図中の四つの検索エンジンは，異なる会社により運営されていたが，2002年末にInktomi社がYahoo!に買収されたのを皮切りに，2003年第一四半期には，Altavista社とAlltheswebを運営するFAST社がOvertune社に買収された．さらに，2003年7月には，Overtune社がYahoo!に買収されている．この結果，Googleを除く三つの検索エンジンは，Yahoo!の配下に入ったことになり，今後の展開が注目される．

4. 全世界Web情報アーカイブ構築

これまでみてきたように，インタネット上には膨大な情報が存在している．このような膨大な情報の中から必要な情報を抽出し有効活用するため，日本国内のWebページを対象に24時間以内で収集することを目指したWebページ収集実験を1998～1999年度に実施すると共に，2003年度からは世界中のWebページを収集・解析するための国家プロジェクトを進めている．

4.1 Webページ分散収集実験

本実験は，筆者を含めたグループが，1998～1999年度に行った実験[4]である．実験は，指数関数的に増大するWebページを効率的に収集するための仕組みを確立することを目的として行われた．

膨大なWebページを高速に収集するためには，並列・分散の考え方が必須となる．一般にWebページを自動収集するためのプログラムをクローラ，またはロボットと呼び，本実験では，WWWロボットをネットワーク上の複数の拠点に配置し，高速収集を目指した．

従来の高速化の主要な方式は，同一のバックボーン下に複数のコンピュータを設置し，複数のWWWロボットを動作させるという方法である．しかし，図6に示すように，①バックボーンとの間に設置されたルータの性能がボトルネックとなったり，②WWWロボットから収集先であるWWWサーバへの経路が遠い場合にそのWWWサーバのWebページ取得に時間がかかる，という問題があり高速化が困難であった．

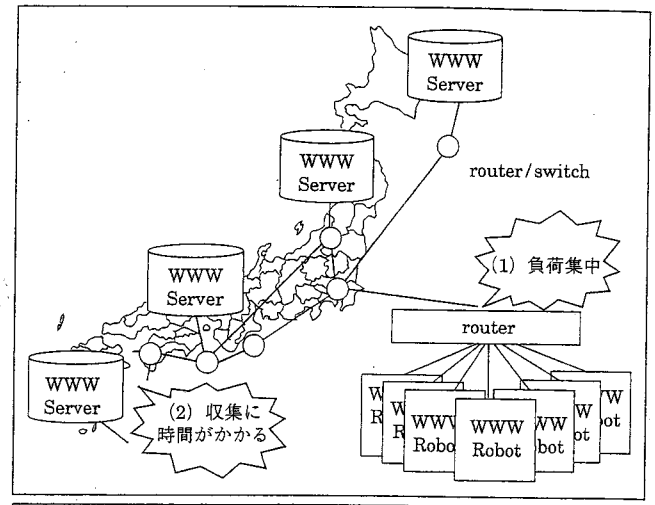


図6 従来の分散型WWWロボットの問題

これらの問題を解決するため，WWWロボットをインタネット上に広域にわたって分散配置すると共に，各WWWロボットが担当するWWWサーバを自動的に決定する手法を，情報処理振興事業協会の独創的情報技術育成事業「インタネット広域分散協調サーチロボット研究開発」として実施した[4]．

本実験には，早稲田大学，京都大学，北陸先端大学院大学，慶應義塾大学，大阪府立大学，日本IBM（株）東京基礎研究所，シャープ（株），電子技術総合研究所の8機関を中心に，外部協力機関を併せて合計33機関（21大学，4インタネットサービスプロバイダ，3企業，3ネットワーク機関，2国立研究所）が参加した（図7）．

本研究が目指した点は，次の2点である．

(1) 広域に分散したWWWロボットを協調動作させ，WWWサーバ上のデータを高速収集する（目標：日本全国のWWWサーバ上のデータを24時間以内に収集）

(2) 検索エンジン運営サイト毎に独立に動かしているWWWロボットを，本研究開発によって開発された広域分散協調サーチロボットに置き換えることにより，WWWサーバの負荷の1/4以上を占めるWWWロボットによる負荷，さらにはインタネットへ与える負荷を大幅に削減する

開発された分散型WWWロボットは，図8に示すように，全体を管理する Public Robot Server Manager（PRSM）と個々のWWWロボットである Public Robot Server（PRS）から構成される．

PRSMは，PRSに対して担当WWWサーバの割当てや，各WWWサーバとPRSとの間のデータ転送速度の計測を指示する．一方，PRSで新規に発見されたWWWサーバやデ

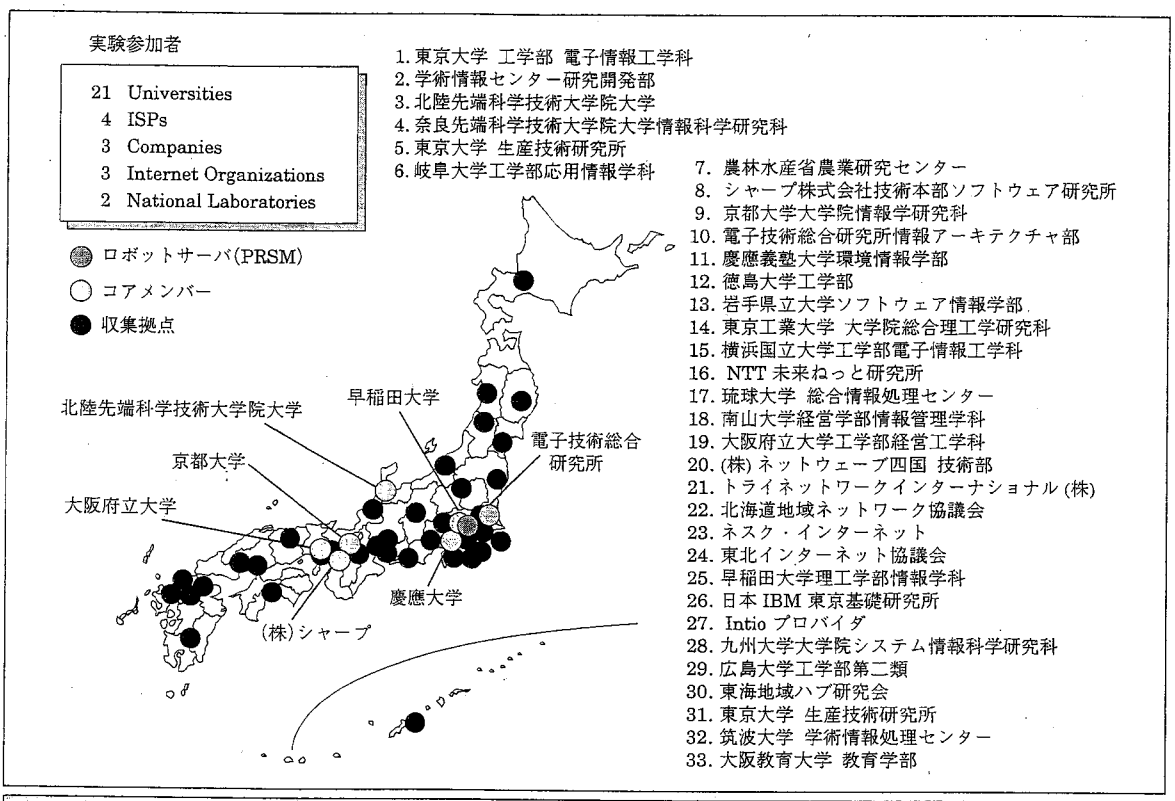
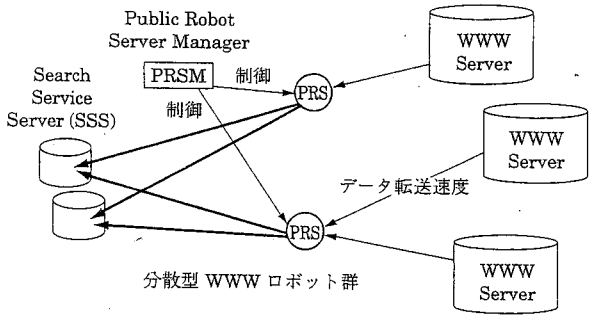
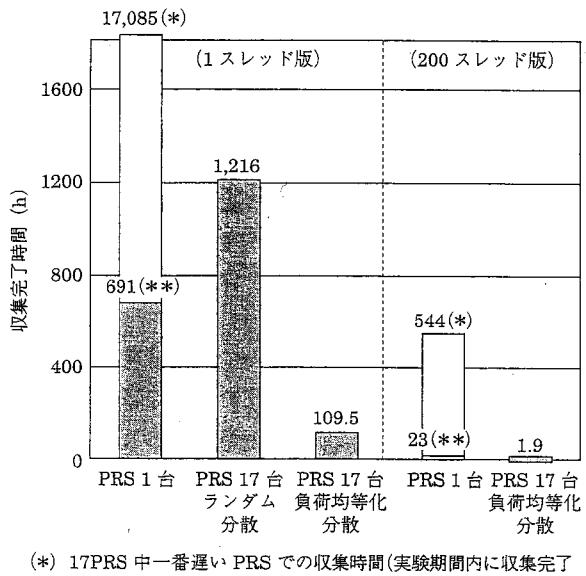


図7 実験参加組織



PRS は複数の WWWServer のデータ収集を担当し(重複設定も可能), 収集した Web ページを圧縮し一つのファイルにした上で, 検索サービスを提供する Search Service Server へ送る. PRS の制御は, PRSM が集中管理する.

図8 実験で使用した分散型WWWロボット



(*) 17PRS 中一番遅い PRS での収集時間(実験期間内に収集完了しなかったため予測値で代用)
(**) 17PRS 中一番速く収集の終了した PRS での収集時間

図9 分散収集による速度向上

ータ転送速度の計測結果は, PRSMに送られる. PRSMは, これらのデータを元に各PRSへのWWWサーバの割当を行う. 具体的には, 得られたデータをもとに, 各WWWサーバを収集するのに必要な時間を予測し, 各WWWロボットが担当するWWWサーバ群の収集完了時間が同じになるように負荷均等化分散を行う[4]. また, 各PRSで収集されたデータは, 最終的に図8中のSearch Service Server (SSS) に再配布され, 検索サービスのための索引作成が行われる.

本プログラムはJavaで記述されており, Javaのスレッドを用いることにより, 一つのPRSから同時に200個のサーバに対して並列アクセスし, 一つのWWWロボットとしても高速化を実現している.

性能評価実験を17台のPRSを用いて行った結果を, 図9に示す. 実験では, JPドメインからランダムに抽出した6,500台のWWWサーバが持つ4,653,140URLを収集対象とした.

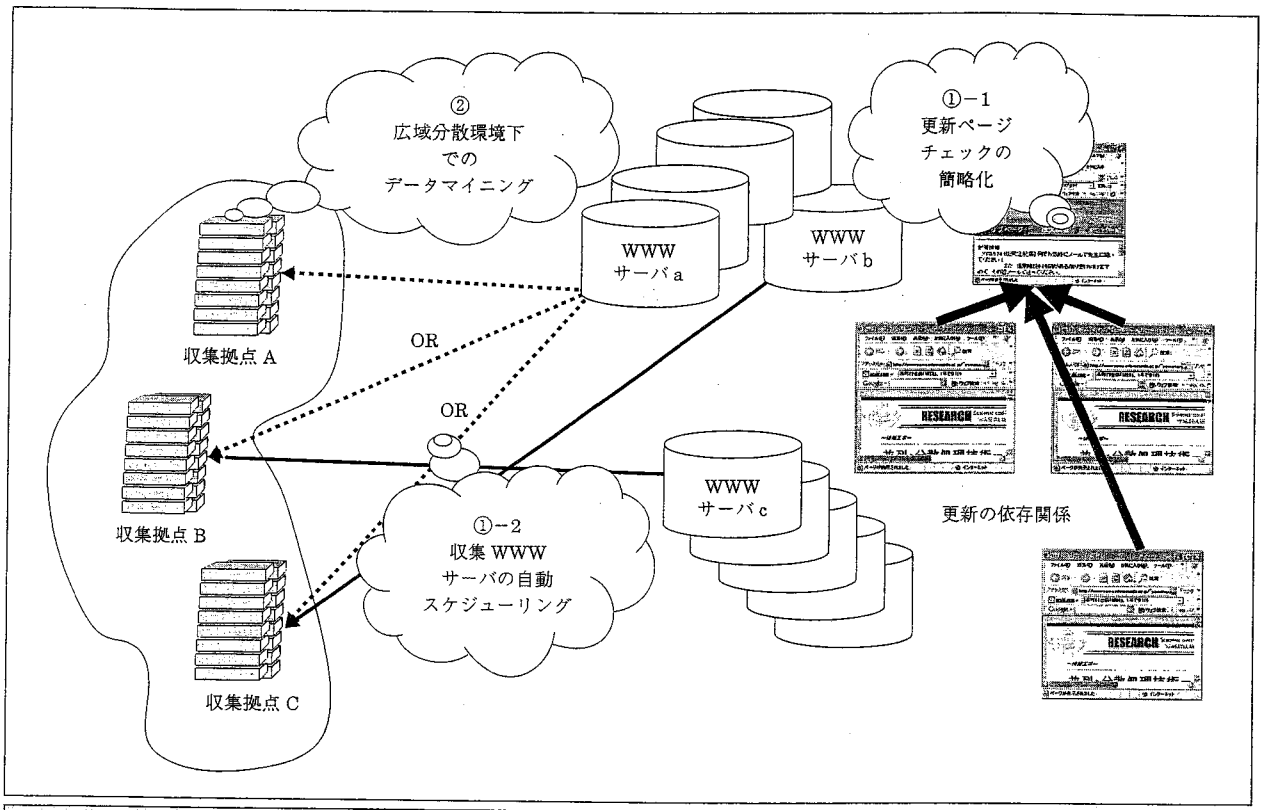


図10 技術的なブレークスルー

図9に示すように，各PRSにランダムに同じ数だけのWWWサーバを分担させるランダム分散では，17台のPRSで分散収集をしても収集が完了するまでの時間が逆に延びてしまう結果が得られている．これは，同一のWWWサーバに対するデータ転送速度が，17台のPRSで大きな差が認められ，一番速いPRSと一番遅いPRSとの差が2倍～710倍，平均67.5倍であることと，各WWWサーバが持つデータ容量に大きな差があることに起因している．

一方，負荷均等化を行った場合には，17台のWWWロボットを用いて 6.3倍（PRS1台で収集した際に最高速であったPRSを基準とした場合）～156倍（PRS1台で収集した際に最低速であったPRSを基準とした場合）の高速化が実現できることがわかった．

さらに，各PRSが200スレッドを同時起動し，200台のWWWサーバに対して並列収集した場合，1台のPRSのみを用いた場合には，最小23時間～最大23日が必要であるが，17台のPRSを用いて分散収集することにより，1.9時間（12～286倍の高速化）で収集可能であることがわかった．

本結果から，高速に収集可能なPRSを基準とした場合には，17台に分散しても，必ずしも17倍の高速化が達成できないことがわかる．これは，例えば，処理能力100の計算機と処理能力50の計算機を使って並列計算した時に最大でも1.5倍にしかならないことと同じであり，ここでの処理能力がPRSとWWWサーバ間のデータ転送速度に対応する．つまり，分散型WWWロボットを使って高速な収集を実現するためには，ネットワーク的に高速収集が可能な地点にPRSを配置することが効率化の点で重要であり，今後の検討課題となった．

4.2 全世界Webアーカイブ構築プロジェクト

2003年度から5ヶ年計画でスタートした文部科学省「e-Society基盤ソフトウェアの総合開発」[5]の「情報の高信頼蓄積・検索技術等の開発」（代表：村岡洋一）においては，最終的に120億規模のデータ*5を対象に，①平均して1ヶ月以内の新しいデータに更新することを可能とするWWWロボットを開発すると共に，②利用者の検索目的に応じて必要となる情報を抽出する知識フィルタリング技術の開発，を目指している．

初年度である2003年度は，前節で紹介した収集実験の経験を踏まえ，国内のネットワークの拠点となる3箇所にWWWロボットを設置し，合計で約50TBのディスクと合計30台のLinuxサーバ機を導入し，10億URLのデータ収集と得られたデータのリンク情報の解析を行っている．また，プロジェクト開始3年後の2005年度までには，合計で約600TBのディスクと合計200台規模のサーバ機を導入することにより，120億規模のWWWデータの収集を実現させ

*5 本開発では，必要となるディスク容量を抑えるため，テキストデータのみを収集の対象としているが，開発される技術は，画像や動画などのマルチメディアデータに対しても有効．

る予定となっている．

本プロジェクトの技術的なポイントを図9に示すと共に，以下で説明する．

(1) 新鮮度平均1ヶ月以内のデータ収集

Webページの平均容量を15kBと仮定すると，120億のWebページは180TBに相当する．仮にバックボーンの帯域として3拠点合計で常に50Mbpsを利用できたとしても，収集に約1年が必要となる．つまり，すべてのWebページを順番に更新した場合，次の更新は1年後になる．

これに対して，本プロジェクトでは，Webページの新鮮度を1ヶ月以内にすることを目標としている．新鮮度とは，インタネット上で実際に公開されているWebページの内容が，収集済のWebページの内容から更新された時点からの経過日数である．

新鮮度を保つための一般的な方法は，従来の検索エンジンで採用されているように，Webページ毎に更新頻度の統計をとり，ページ毎に更新間隔を調整する手法である．例えば，1ヶ月毎に更新チェックを行い，その時点で更新されていれば次のチェックを2週間後に，もし更新されていなければ次のチェックを2ヶ月後に行うという手法である．しかし，120億規模のWebページを対象とした場合，従来手法のみでは，新鮮度1ヶ月以内を達成することは困難である．このため，①Webページの約90%は1ヶ月たっても更新されないという特徴[6]と，②あるWebページが更新される際，その上層のWebページも更新される確率が高いという特徴[6]を用い，一部のWebページのみの更新をチェックすることで，全体の更新を判断できる技術を開発することを目指す．

さらに，異なるバックボーンを持つ3箇所の収集拠点のうち，どこから収集すれば，また，時間帯としていつ収集すれば最も効率的に収集できるかをWWWサーバ毎に自動的に判断しスケジューリングする機能の開発を目指す．

(2) 120億規模のデータを対象としたデータマイニング

本プロジェクトが対象とする解析データの規模は，おおよそ180TBになる．このため，これらの膨大なデータを一箇所に集めて解析することは，事実上不可能である．このため，知識フィルタリングを複数拠点で分散処理する技術を開発する．具体的には，標準的な通信ライブラリーであるMPI (Message Passing Interface) の広域分散環境への対応や耐故障機能の搭載などを行い，従来のデータマイニング手法をクラスタ上で高速化する技術[7]をベースに，広域分散環境下でも適用できる仕組みを開発する．

また，知識フィルタリングの内容としては，利用者が指定する目的に応じて必要となる情報（例：誹謗中傷情報）を自動的に抽出しユーザに提供することを目指す．

5. むすび

「情報」の応用の開拓では，全世界のWeb情報アーカイブ構築を中心に，インタネット上の情報を活用するためのインフラ構築について紹介した．紹介したプロジェクトは，誰もが踏み入れたことのない膨大な情報を対象にした収集・解析を目指しており，これらの基本技術が完成すれば，さまざまな解析に応用できると考えている．

(2003年9月26日受付)

〔文　献〕

1) Steve Lawrence and C.Lee Giles：“Searching the World Wide Web”, Science, 280, 5360, Issue 3, pp.98-100 (Arp. 1998)
2) “Internet Domain Survey”, http://www.isc.org/ds/
3) 山名，近藤：“サーチエンジンGoogle”，情処学誌，42，8，pp.775-780 (Aug. 2001)
4) 山名，森，田村，河野，村岡：“分散型WWWロボットの予備評価と高速化の検討”，日本ソフトウェア科学会，The Third Workshop on Internet Technology (Sep. 2000)
5) “e-Society基盤ソフトウェア総合開発”，http://cif.iis.u-tokyo.ac.jp/e-society/
6) 熊谷，山名：“Webページの更新傾向を踏まえた効率的な収集方法の提案”，第65回情処全大，4ZA-4，pp.3-167-3-168 (Mar. 2003)
7) 岩橋，山名：“FP-growthの並列化による頻出パターン抽出の高速化”，情処研報 (DBS)，2003，71，pp.327-334 (July 2003)

山名　早人（やまな　はやと）　1987年，早稲田大学理工学部通信科卒業．1988年，同大学修士課程修了．1993年，同大学博士課程修了．1993年～2000年，電子技術総合研究所．1996年～1997年通商産業省機械情報産業局電子機器課．2000年，早稲田大学理工学部助教授，現在に至る．並列・分散処理技術，情報検索技術に従事．工学博士．